OPEN ACCESS

KING'S College LONDON

King's Research Portal

DOI:
10.1034/j.1600-0447.2002.02356.x

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Citation for published version (APA):
Slade, M., Cahill, S., Kelsey, W., Powell, R., & Strathdee, G. (2002). Threshold 2: The reliability, validity and sensitivity to change of the Threshold Assessment Grid (TAG). *Acta Psychiatrica Scandinavica*, *106*(6), 453 - 460. https://doi.org/10.1034/j.1600-0447.2002.02356.x

Citing this paper
Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights
Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

•Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
•You may not further distribute the material or use it for any profit-making activity or commercial gain
•You may freely distribute the URL identifying the publication in the Research Portal

Take down policy
If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

Slade M, Cahill S, Kelsey W, Powell R, Strathdee G (2002) *Threshold 2: The reliability, validity and sensitivity to change of the Threshold Assessment Grid (TAG),* Acta Psychiatrica Scandinavica, **106**, 453-460.

Threshold 2: The reliability, validity and sensitivity to change of the Threshold Assessment Grid (TAG)

Running head: The psychometric properties of the TAG

M Slade
S Cahill
W Kelsey
R Powell
G Strathdee

Address for correspondence:
Dr Mike Slade
MRC Clinician Scientist Fellow
Health Services Research Department
Institute of Psychiatry
De Crespigny Park
Denmark Hill
London
SE5 8AF
England

Tel: 020 7848 0714
Fax: 020 7277 1462
Email: m.slade@iop.kcl.ac.uk

15 May 2002

Threshold 2: The reliability, validity and sensitivity to change of the Threshold Assessment Grid (TAG)

Running head: The psychometric properties of the TAG

Abstract

Objective

This study investigated the psychometric properties of the Threshold Assessment Grid (TAG), a new assessment of the severity of mental health problems.

Method

605 patients were recruited from ten mental health adult and elderly services in London, England. TAG ratings and other standardised definitions of severe mental illness were completed by referrers. TAG, GAF, CANSAS and HoNOS ratings were completed by mental health service staff. Construct validation on extreme groups was investigated.

Results

Construct and concurrent validity were good. Referrer TAG scores predicted mental health team view of referral suitability, but not whether assessments were offered. Test-retest reliability was good, inter-rater reliability ranged from good to poor in different domains (but adequate for total TAG score), internal consistency was appropriate. Sensitivity to change requires further investigation.

Conclusion

The Threshold Assessment Grid can be recommended for use by all agencies when making referrals to mental health services.

Key words

Psychometrics, Mental Health, Primary Health Care, Referral

Introduction

Specialist mental health services are a scarce resource, and need to be effectively targeted towards people with more severe and enduring mental health problems. This apparently simple goal has proved difficult to achieve, yet the need to match referrals with available services is becoming more pressing: referral rates from primary to secondary care for mental health problems have increased by a factor of 4.5 from 1971 to 1997 [1]. The difficulty in getting the right patients referred to specialist mental health services is highlighted by the mismatch between this increasing rate of referral and the patient preference for primary-care level talking therapy over medication or referral to a mental health professional [2].

Maximising the proportion of appropriate referrals requires at least two developments: shared agreement about who the severely mentally ill are, and a currency for communication between primary care services (such as health and social services) and specialist mental health services. One problem hampering the development of community care is the lack of a shared definition of severe mental illness [3]. In England a research team – the Threshold Programme – have been working on this problem since 1994. Following an international survey [4, 5], the need to develop a new assessment schedule to address the lack of consensus between agencies was identified. Accordingly, innovative consensus-testing techniques (search workshops and Delphi Consultation) were used from 1997 to 1998 to develop an assessment that is acceptable to all relevant stake-holders, including mental health service users and carers, primary and secondary health services, social services, housing services, care commissioners, and policy-makers [6]. The resulting assessment – the Threshold Assessment Grid (TAG) – is intended for routine multi-agency use, when making a referral to specialist mental health

services. The development process, unlike the traditional method of developing psychiatric assessment tools, ensured the external validity of the TAG, since its content was derived directly from the views of the full range of relevant stake-holders.

The TAG is intended to improve the process of making referrals to specialist mental health services, both by supporting referrers in assessing mental health problems, and informing decision-making about when to refer the patient on to a specialist mental health team. The aim of this study was to complete the next stage in the systematic development of the TAG, by investigating its reliability, internal validity and sensitivity to change when it is used in routine (*i.e.* non-research) mental health services.

Material and Method

Assessments

The development of the TAG has been reported previously [6]. The TAG is a 1-page assessment of the severity of a person's mental health problems. It is completed by making one tick to indicate level of severity in each of 7 domains: (i) intentional self-harm; (ii) unintentional self-harm; (iii) risk from others; (iv) risk to others; (v) survival needs/disabilities; (vi) psychological needs/disabilities; and (vii) social needs/disabilities. The scale is "None", "Mild", "Moderate" and "Severe" (4-point scale, ranging from 0-3) for domains (ii), (iii), (vi) and (vii), with an extra "Very Severe" domain (score 4) possible for the remaining 3 domains (which may require immediate action). It is completed by making 7 ticks. A second page of the TAG gives guidance for rating each domain, based on research

evidence and best clinical practice. The TAG is intended for use without training by any health professional, although a third page includes detailed instructions for using the TAG.

Scores for the TAG can be presented in any of three ways. Seven individual Domain Scores can be presented in order, to provide a profile that can be used for longitudinal monitoring. A single Total TAG score can be calculated by summing the domain scores. Finally, a threshold can be used to create categories, such as "severely mentally ill" versus "not severely mentally ill", or low, medium and high priority bands. Only the first two of these scores are reported here.

Two types of staff participated in this study – referrers to and staff within secondary mental health services. As well as the Threshold Assessment Grid (TAG), referrers completed two categorical definitions of severe mental illness (SMI). The first (**SMI 1**) considers level of social functioning in the context of severe and persistent mental illness, and categorises patients into SMI or non-SMI [7]. The second (**SMI 2**) considers diagnosis, disability and duration, and categorises patients into low, medium or high support needs groups [8]. Although no consensus exists about a gold standard [5], these definitions were chosen as research-based assessments from primary (SMI 1) and secondary (SMI 2) care perspectives.

Mental health service staff completed the TAG, the Global Assessment of Functioning (GAF) [9], the Health of the Nation Outcome Scale (HoNOS) [10], the Camberwell Assessment of Need Short Appraisal Schedule (CANSAS) [11], and an item on social vulnerability – question 4 from the 'Social circumstances' section of the CORE assessment [12]. (A component of this

questionnaire was used because no specific assessment on social vulnerability could be located). The order of administration was rotated to control for response bias.

The GAF is a staff-rated global measure of symptomatology and social functioning, with a scale ranging from 1 (worst) to 99 (best). The HoNOS is a staff-rated, 12-item assessment of social disability, with each item rated from 0 (no difficulties) to 4 (severe to very severe), with 9 indicating not known. The social vulnerability question uses the same ratings as HoNOS. The CANSAS is a staff--rated 22-item assessment of health and social needs, with each item rated either 0 (no need), 1 (met need), 2 (unmet need) or 9 (not known).

Setting and Subjects

Ten community-based routine (*i.e.* long-term National Health Service-funded) mental health services in London were included in the study, which took place between June 1999 and September 2000. The ten sites comprised eight adult, one adult day care and one elderly mental health team. In terms of deprivation levels, the Mental Illness Needs Index [13] scores (mean 100, higher score = more deprived) for the three inner London sites ranged from 120 to 124, for the five outer London sites from 98 to 121, and for the two suburban sites from 107 to 108. Referrals for 605 patients were included. For the adult sample (n=544) the mean age was 37.7 (s.d. 11.7), 246 (45%) were male, and the most common clinical diagnoses were depression (42%), psychosis (15%), anxiety (7%) substance misuse (6%) and unspecified (14%). For the elderly sample (n=61) the mean age was 78.3 (s.d. 7.1), 27 (44%) were male, and the most common clinical diagnoses were dementia (47%), depression (23%), psychosis (7%) and unspecified (13%).

Methods

For each service, recent referrals were retrospectively audited to identify the most frequent referrers. Letters were sent to these referrers and other local voluntary sector organisations describing the study and asking for their participation. Where possible the letter came from the local clinical or medical director. Training in the use of TAG, GAF, HoNOS and CANSAS was offered to all clinical members of the secondary mental health service.

The next 60 consecutive referrals from all referrers (plus self-referrals or informal carer referrals) to the service were included in the study. For each referral, the referrer was contacted within 1 working day of the referral being discussed, with a request to complete a TAG (the **referrer TAG**) and SMI 1 and SMI 2 (for concurrent validity). Date of birth and clinical diagnosis were recorded from the referral. The pathway through care was then tracked for the patient, from whether the team decided to offer an assessment appointment (for predictive validity), up to and including the second clinical contact with the first team member to assess (for test-retest reliability) and also the first contact with any subsequent member of the team (for inter-rater reliability). After the first clinical contact, the mental health service clinician completed the **initial TAG**, CANSAS, HoNOS, GAF, and the CORE question (for concurrent validity). After their second contact they were asked again to complete the TAG (the **test-retest TAG**). When a second clinician had contact with the patient, either during a joint assessment or subsequently, they were also asked to complete a TAG (the **inter-rater TAG**). To increase the reliability sample, 12 extra inter-rater reliability ratings and 23 extra test-retest ratings were gathered from a convenience sample of other

patients seen by the clinicians involved in the study. All assessments were sent directly to the researcher, and the clinical team was blind to the TAG and other assessments by the referrer when they made their decision about responding to the referral. Returned assessments which were incorrectly completed or blank were ignored, which involved 34 HoNOS (11%), 25 GAF (8%) and 23 CANSAS (7%) assessments. Nine (2%) partly-completed TAGs from referrers and three (1%) from mental health teams were either pro-rated (where 5 or 6 domains were completed) or assumed to have a 0 rating for missing domains.

Three measures of predictive validity were gathered. First, the decision as to whether or not to offer an assessment was recorded for all patients. Second, after the initial assessment, the assessor was asked to identify whether the patient was a suitable referral for the team, and if not to identify whether the person was either too severely or insufficiently severely mentally ill to warrant mental health service involvement, or more suitable for another agency. This was monitored for all referrals after January 2000. Finally, as a proxy measure for those patients who were accepted by the service, those patients seen twice or more were considered.

In a separate exercise, construct validation on extreme groups was investigated by the development of 20 vignettes, which were previously identified by two expert raters as describing ten SMI (at least one HoNOS item score of 4 or two scores of 3 on at least 2 different items on items 1-10 excluding item 5 or a total score of 12 PLUS GAF score of 48 or less) and 10 non-SMI patients, with cut-off points chosen based on a policy definition of severe mental illness [14]. Ten experienced clinicians (defined as profesisonals who are more than 5 years post-qualification) were then asked to rate each vignette using the TAG.

Analysis

For concurrent validity analysis, CANSAS item scores for met and unmet needs were recoded into a single category, and differences in TAG ratings between groups identified by SMI 1 were investigated using an independent samples t-test, which was also used for predictive validity comparing TAG ratings for the various predictive factors. ANOVA with a Bonferroni adjustment for the pairwise comparisons between adjacent groups was used for the three SMI 2 groups. Stability (test-retest reliability, inter-rater reliability, and sensitivity to change) analyses was investigated using average measure intraclass correlations. Cronbach's Alpha was used for internal consistency. Construct validation was tested using paired samples t-tests comparing SMI and non-SMI vignettes from the same rater using both weighted (summing the domain scores with 0 for each "None" domain, 1 for "Mild", 3 for "Moderate", 5 for "Severe" and 7 for "Very severe" – maximum score 41) and unweighted (summing domains scoring 0 for "None", 1 for "Mild", 2 for "Moderate", 3 for "Severe" or 4 for "Very severe" – maximum score 24) TAG total scores. Following logistic regression on the weighted and the unweighted totals, the TAG weighting was investigated using ROC curve analysis – a graphical representation of the trade-off between the false negative and false positive rates for every possible cut-off, in which the area under the curve represents the accuracy of the test (in this case whether the TAG score discriminates between SMI and non-SMI). To take account of dependence due to multiple assessments by the same person, this analysis was clustered by assessor. To take account of skewness, the findings were confirmed using square-root transformed data (not presented). All analysis was undertaken using SPSS 8.0 and Stata 6.0.

Results

605 Patients were included in the study (range 60-62 per site). 600 Patients were referred by 394 individual professionals, comprising 249 General Practitioners (63%), 78 psychiatrists (20%), 24 psychiatric nurses (6%), 21 (5%) care managers, 6 liaison mental health team staff, and 16 others (comprising health visitors, parole officers, physicians, psychologists, housing officers, alcohol workers and voluntary sector workers). Referrers referred between 1 and 7 people, with the exception of two psychiatrists (10 and 13 referrals) referring to the day care service. In addition, 5 patients were self-referrals or referred by family members (for whom referrer data were not collected). 445 (74%) Referrers completed TAGs, of whom 399 (90%) completed SMI 1 (with 205 (51%) patients identified as severely mentally ill) and 421 (95%) completed SMI 2 (with 86 (20%) patients assessed as needing high support, 178 (42%) as needing medium support, and 157 (37%) as needing low support). There was no evidence that the 155 patients whose referrers did not complete the TAG differed from the 445 patients whose referrers did complete the TAG – 88 (55%) were male, their mean age was 36.4 years for the adult group and 78.7 for the elderly group, and the most common clinical diagnoses were depression (36%), psychosis (16%), anxiety (7%), substance misues (5%), physical illness (5%), dementia (4%) and unspecified (7%).

350 Patients were seen for initial assessment, following which TAG data were collected for 308 (88%). 101 Mental health service staff completed initial TAGs, comprising 39 psychiatric nurses, 41 psychiatrists, 11 clinical psychologists, 7 occupational therapists, 1 care manager and 1 art therapist. The referrer and initial TAG assessments are shown in Table 1.

Insert Table 1 here

308 Mental health team TAGs were completed, along with 285 CANSAS ratings (mean 3.2 (s.d. 2.6) met needs, 4.0 (3.0) unmet needs, 7.2 (3.7) total needs), 283 GAF ratings (mean 59.2, s.d. 13.7), 274 HoNOS ratings (mean 9.8, s.d. 5.1) and 274 social vulnerability questions (mean 1.1, s.d. 1.2).

Construct validity

The vignettes were rated by 2 general practitioners, 3 consultant psychiatrists, 2 clinical psychologists, 1 occupational therapist, 1 liaison mental health nurse, and 1 community psychiatric nurse. The total TAG score for the non-SMI group (n=10 x 10=100; *i.e.* 10 vignettes, 10 clinicians) was 3.4 (s.d. 2.0) and for the SMI group (n=100) it was 15.0 (s.d. 3.0), with a paired mean difference of 11.7 (95%C.I. 10.9-12.4) ($t=29.5$, $df=99$, $p<0.001$)[a].

Concurrent validity

The mean domain-specific and total referrer TAG ratings for patients in each SMI sub-group are shown in Table 1. There were significant differences between total TAG scores for groups identified by SMI 1 groups ($t=6.3$, $df=396$, $p<0.001$). The three SMI 2 groups had differing total TAGs (4.92 vs 6.54 vs 9.44: $F=53.4$, $df=2,417$, $p<0.001$), with increasing total TAGs across the ordered groups.

[a] The option of using a different scoring scale for producing the TAG score (with a heavier weighting for more severe ratings) was also investigated, but discounted because the areas under the ROC curves for unweighted and weighted TAGs were very similar (0.999 versus 0.998 respectively).

The correlation of initial TAG total with GAF total was -.65 ($p<0.001$), with HoNOS total was .71 ($p<0.001$), and with CANSAS total was .53 ($p<0.001$). The correlation between TAG domain 1 (intentional self-harm) and HoNOS item 2 (suicidal thoughts or behaviour or non-accidental self-injury) was .71 ($p<0.001$), between TAG domain 3 (risk from others) and the social vulnerability question was .45 ($p<0.001$), between TAG domain 4 (risk to others) and HoNOS item 1 (overactive, aggressive, disruptive or agitated behaviour) was .52 ($p<0.001$), between TAG domain 6 (psychological needs/disabilities) and GAF symptoms was -.56 ($p<0.001$), and between TAG domain 7 (social needs/disabilities) and (a) GAF disability was -.52 ($p<0.001$) and (b) HoNOS item 9 (social relationships) was .58 ($p<0.001$).

There was a significant difference between TAG domain 1 (intentional self-harm) scores for patients having or not having a CANSAS "safety to self" need (.21 vs 1.20, t=13.4, df=306, 95%CI for the difference .85 – 1.14), between TAG domain 2 (unintentional self-harm) scores for CANSAS "self care" (.42 vs .84, t=4.0, df=306, 95%CI .21 – .61), between TAG domain 4 (risk to others) scores for CANSAS "safety to others" (.14 vs 1.38, t=13.5, df=306, 95%CI 1.06 – 1.43), between TAG domain 5 (survival needs/disabilities) scores for CANSAS "physical health" (.56 vs .76, t=2.2, df=306, 95%CI .02 – .39), and between TAG domain 7 (social needs/disabilities) scores for CANSAS "company" (0.99 vs 1.45, t=4.9, df=306, 95%CI .27 – .64). All differences were in the predicted direction.

Predictive validity

Three measures of predictive validity were used for the 600 patients referred by professionals. 476 (79%) were offered an assessment. 168 (35%) of these were seen twice. Following initial

assessment, the staff member thought 103 (79%) of 131 patients were suitable referrals. The total referrer TAG did not predict whether the patient was seen (6.05 vs 6.53) or whether they were seen twice (6.29 vs 6.79). It did predict the rating of suitability (5.20 vs 6.84, t=2.1, df=94, p=0.04). Initial TAGs (by mental health staff) were significantly different in the predicted direction for those seen twice versus not seen twice (4.55 vs 5.61, t=3.0, df=304, p=0.004) and for those rated as suitable (3.19 vs 5.57, t=3.3, df=115, p=0.001).

Stability

142 Repeat TAGs and 78 inter-rater TAGs were completed, including 57 inter-rater assessments done on the same day. Test-retest reliability was investigated for the 65 patients with repeat TAGs no more than 14 days after the initial TAG. Sensitivity to change was investigated for the 44 patients with a repeat TAG more than 29 days after the initial TAG. Inter-rater reliability was considered for the 62 patients seen no more than 14 days apart by the two raters. Reliability and sensitivity to change analyses are shown in Table 2.

Insert Table 2 here

Sensitivity to change was further investigated by comparing the mean difference scores for the patients seen less than 15 days apart (6.55 vs 5.6, difference=0.95, t=3.5, df=64, p=0.001) and those seen more than 29 days apart (5.18 vs 4.16, difference=1.02, t=3.0, df=43, p=0.004). The internal consistency for the total referrer TAG assessments (n=445) was 0.73 and for the initial TAGs (n=307) was 0.70.

Discussion

This study investigated the internal consistency, construct validity, concurrent validity, predictive validity, and test-retest and inter-rater reliability of the Threshold Assessment Grid (TAG), an assessment of the severity of mental health problems. The prospective cohort study took place in typical mental health services, and the findings are therefore likely to be indicative of the psychometric properties of the TAG when used routinely. The main findings were that the TAG had good construct validity, good concurrent validity, referrer TAGs predicted whether the referral was seen as suitable for the service but not whether the patient was offered one or more than one appointment, mixed inter-rater reliability, good test-retest reliability, and an appropriate level of internal consistency. Return rates of 74% for referrers and 88% for mental health service staff were obtained. Sensitivity to change requires further investigation.

The TAG strongly discriminated between divergent groups, indicating that (as in real life) any lack of discrimination is likely to occur in the middle group of patients, about whom clinical judgement may in any event differ. For example, there is evidence of differences between professions in the importance attached to different aspects of assessment [15] or types of treatment [16].

The simpler weighting of TAG domains (1 ‘point’ for each increase in severity) was chosen, on the basis that the ability to discriminate between vignette groups was equally high whichever rating was used, and because of the intuitive appeal of a simple rating system. However, it could be argued that the weighted TAG fits the clinical situation better, since

many people have mild problems in several domains of life, and the 'clinical threshold' is crossed when these become moderate problems. The question of how to weight the TAG might have been better addressed using vignettes chosen to span the spectrum of severity, though such an approach would require decisions about the trade-off between sensitivity and specificity, which might be problematic given that the intention is that the TAG be a nationally-applicable assessment, so any decision about the balance between sensitivity and specificity may not be appropriate for all settings.

A strength of this study is its representativeness of people referred to mental health services across London. In particular, the diagnostic heterogeneity is more representative of clinical practice than a diagnosis-based inclusion criterion. In research trials, it is common to equate 'severe mental illness' with 'psychosis', but this substantially reduces the external validity (*i.e.* clinical relevance) of such studies, since diagnosis does not predict need for service [17, 18]. The use of TAG within rural services was not investigated.

There was strong evidence for the internal validity of the TAG. The concurrent validity data indicated that, with the exception of referrer rating of domain 3 (Risk from others) for SMI 2 (which features in neither standardised definition used), all expected associations were present, in the expected direction. This implies that the TAG is assessing in a similar way to other standardised assessments.

Predictive validity was more complex to investigate, because of difficulties in identifying what a TAG score should predict. The study was originally intended to monitor what teams

did (*i.e.* offering one or more than one appointment), rather than their stated view, to reduce bias due to pre-existing opinions about the referrer. However, it became clear that complex factors influenced how teams responded to referrals, and that these factors differed between sites. For example, in this study only 7 of the 10 teams had written referral protocols, of which 4 were shared with referrers and none were jointly developed with referrers. Some services had an ethos that all referred patients should be assessed, and others required eligibility criteria to be met before an assessment was offered. It was the researchers' impression that being offered an assessment was subject to many more influences than simply the team perception about the severity of mental health problems. Therefore, a specific question was introduced mid-way through the study to more directly ascertain the view of suitability following their initial assessment, and the referrer TAG predicted the response. The initial TAG also predicted how the team both viewed and responded to the referral. Overall, some evidence for predictive validity was found.

There was strong evidence for test-retest reliability in individual domains and total TAG scores – Streiner and Norman recommend a minimum reliability of 0.75 for classification tests (in this case, whether the person's mental health problems are sufficiently severe to warrant specialist mental health service involvement) [19]. The evidence for inter-rater reliability was much weaker, with strong inter-rater reliability present for only one domain (intentional self-harm), but acceptable reliability for TAG total scores between raters. Assessment of risk to others and social needs / disabilities had particularly poor agreement between raters, and with respect to social needs it may be relevant that most of the assessors were health rather than social service professionals. Experience with HoNOS indicates that

the development of routine clinical outcome measures with acceptable psychometric properties is difficult [20], and in particular inter-rater agreement is known to be problematic [21]. It may be that inter-rater reliability for such assessments will increase only when training and the opportunity for practice in using standardised assessments is prioritised in routine clinical practice [22], or that a new generation of 'feasible' assessments need to be developed explicitly for routine clinical use [23]. The feasibility of the TAG has been reported elsewhere [24].

The intra-class correlations and the differences between mean scores were both higher for the group whose repeat assessment was more than 29 days later than for those assessed fewer than 15 days apart, providing some evidence that the TAG is not insensitive to change. However, no comparator "gold standard" was used to identify those patients whose mental health problem severity had altered, so further work is needed to investigate the sensitivity to change in TAG ratings for patients over time. It should be noted, however, that the TAG is not designed to replace careful clinical assessment (which should identify even small changes), but rather is intended to be used as an overall assessment of severity, suitable in particular for use between agencies.

The internal consistency for two separate groups of ratings (referrer and mental health service staff) were both in the range indicated by Streiner and Norman [19], implying the TAG has adequate internal consistency whilst avoiding redundancy in its items.

This study differs from traditional psychometric studies in a number of respects, which may be regarded as weaknesses. Firstly, the degree of characterisation of the study sample was

poorer. Since the properties of the TAG were being tested in the context of routine clinical practice, a research diagnosis was not possible. The heterogeneity of the primary clinical diagnosis (not all of which were made by medically trained staff) is representative of the information routinely provided in referrals to mental health services, and therefore probably lacks strong reliability. Secondly, TAG assessments were only available for patients actually seen, which (assuming that systematic decision-making underpins the decision to assess) will introduce a systematic bias into the TAG data which was collected. Both these issues could be addressed by either controlling the clinical contact or by the use of external researchers, but this rigour would be at the expense of external validity of the findings.

Finally, and perhaps most significantly, every assessment in this study was completed by a referrer or mental health practitioner, rather than (as is frequently the case in psychometric studies) a researcher. This means that the data collected are likely to have a lower return rate and be of a poorer quality than is possible with a dedicated researcher. In other words, the resulting psychometric properties of the instrument are likely to be compromised when compared with what would be obtained in a researcher-based study. However, since the TAG is intended to be used in routine clinical practice, it is hoped that the data resulting from this study may be of more relevance.

Three other weaknesses can be identified. First, no data were available on those patients whom referrers had decided not to refer, either due to non-detection [25] or judging referral as unnecessary. Therefore no conclusions can be drawn as to whether using TAG improves recognition of patients who would benefit from referral to mental health services. This

indicates the need for research examining what factors impact on referrers deciding to refer, and whether referrals are improved when using the TAG. Second, the findings need to be replicated in rural settings, and the inclusion of only one elderly service may have limited generalisibility. Third, the processes underpinning the response by mental health teams to referrals needs to be disaggregated. TAG is intended to identify people with more severe mental health problems, but it was clear that other factors (such as the team's relationship with the referrer and their current caseload) were also impacting on the decision to assess.

In conclusion, the TAG assesses the severity of mental health problems, and is the result of a fresh approach to the development of assessments intended for routine clinical practice. As part of this strategy it is included as an appendix in the development paper [6], and a downloadable version and on-line training are available from www.iop.kcl.ac.uk/prism/tag. TAG is intended to be more suitable for multi-agency use than GAF, and more appropriate for primary care use than HoNOS. It was developed using consensus and search techniques to maximise its external validity, and only subsequently in this study have its other psychometric properties been tested. On the basis of the findings from this multi-site prospective study, it meets most criteria for validity well. In particular, the total referrer TAG predicted whether the referral was assessed as suitable by the mental health service, indicating that TAG may contribute to inter-agency communication and the development of primary/secondary shared care protocols. It suffers from the same difficulties as other routine assessments with respect to inter-rater reliability, but its test-retest reliability and internal consistency are adequate.

Given the paucity of evidence-based approaches to mental health referral protocols, the TAG can be tentatively recommended as a currency for the local negotiation of referral thresholds between primary care health and other agencies and specialist mental health services. It is intended to augment, rather than replace, referral letters, and this study provides evidence that its use in this way will improve communication between referrer and mental health service. Using the TAG will be a step in the direction of evidence-based mental health services, which are referred people with more severe mental health problems, provide more help to those with more needs, and have an overall goal of improving quality of life [26].

Acknowledgements

We are grateful to all the clinical and administrative staff at the ten sites involved in the study. Agitha Valiakalayil provided valuable assistance with data collection, and further assistance was provided by Margot Croft, Alex Dionysius, Dr Hilary Guite, Jimmy Kinsella, Dr Morven Leese, Mauricio Moreno, and Dr Christine Stone. This study was funded by North Thames Responsive Funding Programme (RFG549). The views expressed in the publication are those of the authors and not necessarily those of the NHS Executive or the Department of Health.

References

1 VERHAAK PF, VAN DE LISDONK EH, BOR JH, HUTSCHEMAEKERS GJ. GPs' referral to mental health care during the past 25 years. *Br J Gen Pract* 2000;**50**:307-8.

2 BRODY DS, KHALIQ AA, THOMPSON TL. Patients' perspectives on the management of emotional distress in primary care settings. *J Gen Intern Med* 1997;**12**:403-6.

3 House of Commons Select Committee. *Better off in the community? The care of people who are seriously mentally ill*. HMSO: London, 1994.

4 POWELL R, SLADE M. (1996). Defining Severe Mental Illness. In: THORNICROFT G, STRATHDEE G ed. Commissioning Mental Health Services. London: HMSO, 1996: 13-27.

5 SLADE M, POWELL R, STRATHDEE G. (1997). Current approaches to identifying the severely mentally ill. *Soc Psychiatr Psychiatric Epidemiol* 1997;**32**:177-184.

6 SLADE M, POWELL R, ROSEN A, STRATHDEE G. Threshold Assessment Grid (TAG): the development of a valid and brief scale to assess the severity of mental illness. *Soc Psychiatr Psychiatric Epidemiol* 2000;**35**:78-85.

7 Department of Health. *Health of the Nation Key Area Handbook, 2nd Edition*. HMSO: London, 1994.

8 KENDRICK T, BURNS T, FREELING P, SIBBALD B. Provision of care to general practice patients with long-term mental illness: a survey in 16 practices. *Br J Gen Pract* 1994;**44**:301-305.

9 American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (4th edn) (DSM-IV)*. Washington, DC: APA: 1994.

10 WING J, BEEVOR A, CURTIS RH, PARK SB, HADDEN S, BURNS A. Health of the Nation Outcome Scales (HoNOS). *Brit J Psychiatr* 1998;**172**:11-18.

11 SLADE M, THORNICROFT G, LOFTUS L, PHELAN M, WYKES T. The Camberwell Assessment of Need. London: Gaskell, 1999.

12 CLIFFORD M. The FACE Recording and Measurement System. *Bull Menninger Clin* 1999;**63**:305-331.

13 GLOVER GR, ROBIN E, EMAMI J, ARABSCHEIBANI GR. A needs index for mental health care. *Soc Psychiatr Psychiatric Epidemiol* 1998;**33**:89-96.

14 Department of Health *Mental Health National Service Framework*. London: HMSO, 1999.

15 SLADE M, ROSEN A, SHANKAR R. Multidisciplinary mental health teams. *Int J Soc Psychiatry* 1995;**41**:180-189.

16 JORM AF, KORTEN AE, JACOMB PA, RODGERS B, POLLITT P. Beliefs about the helpfulness of interventions for mental disorders: a comparison of general practitioners, psychiatrists and clinical psychologists. *Aust N Z J Psychiatry* 1997;**31**:844-851.

17 MCCRONE P, STRATHDEE G. Needs not diagnosis: towards a more rational approach to community mental health resourcing in Britain. *Int J Soc Psychiatry* 1994;**40**:79-86.

18 BEDIRHAN ÜSTÜN T. Unmet need for management of mental disorders in primary care. In: ANDREWS G, HENDERSON S. ed. Unmet need in psychiatry. Cambridge: Cambridge University Press, 2000: 157-171.

19 STREINER D, NORMAN L. *Health measurement scales, 2nd edition*. Oxford: Oxford University Press, 1995.

20 BEBBINGTON P, BRUGHA T, HILL T, MARSDEN L, WINDOW S. Validation of the Health of the Nation Outcome Scales. *Brit J Psychiatr* 1999;**174**:389-394.

21 ORRELL M, YARD P, HANDYSIDES J, SCHAPIRA R. Validity and reliability of the Health of the Nation Outcome Scales in psychiatric patients in the community. *Brit J Psychiatr* 1999;**174**:409-412.

22 CRIPPA JAS, SANCHES RF, HALLAK JEC, LOUREIRO SR, ZUARDI AW. A structured interview guide increases Brief Psychiatric Rating Scale reliability in raters with low clinical experience. *Acta Psychiatr Scand* 2001;**103**:465-70.

23 SLADE M, THORNICROFT G, GLOVER G. The feasibility of routine outcome measures in mental health. *Social Psychiatr Psychiatric Epidemiol* 1999;**34**:243-9.

24 SLADE M, CAHILL C, KELSEY W, POWELL R, STRATHDEE G, VALIAKALAYIL A. Threshold 3: The feasibility of the Threshold Assessment Grid (TAG) for routine assessment of the severity of mental health problems. *Social Psychiatr Psychiatric Epidemiol* 2001;**36**:516-21.

25 KARLSSON H, JOUKAMAA M, LEHTINEN V. Differences between patients with identified and not identified psychiatric disorders in primary care. *Acta Psychiatr Scand* 2000;**102**:354-8.

26 SLADE M, LEESE M, TAYLOR R, THORNICROFT G. The association between needs and quality of life in an epidemiologically representative sample of people with psychosis. *Acta Psychiatr Scand* 1999;**100**:149-57.

Table 1: mean TAG ratings by referrers and mental health service staff, and referrer TAG ratings for SMI sub-groups

Mean total TAG score	Possible Range	**Referrers**	**Mental health team**[b]	**referrer TAG vs SMI 1**		**referrer TAG vs SMI 2**		
TAG domain		(n=445)	(n=308)	**No** (n=195)	**Yes** (n=204)	**Low** (n=157)	**Medium** (n=178)	**High** (n=86)
1. Intentional self-harm	0-4	0.73	0.54	0.64	0.85*	0.64ns	0.67	1.08
2. Unintentional self-harm	0-3	0.82	0.53	0.65	1.03	0.58ns	0.78	1.47
3. Risk from others	0-3	0.54	0.44	0.51	0.72ns	0.45ns	0.48	0.85
4. Risk to others	0-4	0.45	0.28	0.34	0.60	0.24	0.50	0.80*
5. Survival needs / disabilities	0-4	0.76	0.64	0.54	1.02	0.41	0.83	1.27
6. Psychological needs / disabilities	0-3	1.65	1.44	1.48	1.85	1.37	1.78	2.05
7. Social needs / disabilities	0-3	1.47	1.28	1.29	1.69	1.24*	1.51	1.92

[b] The referrer TAG and mental health team TAG scores relate to different patient sub-groups, so are not directly comparable.

	Total	0-24	6.43	5.13	**5.45**	**7.64**	**4.92**	**6.54**	**9.44**

All inter-group differences for SMI sub-groups significant at $p<0.01$, except * $p<0.05$, ns=non-significant

All SMI2 low versus high group differences significant at $p<0.001$

Table 2: Reliability and sensitivity to change

Intra-class correlation (95% C.I.)	**Inter-rater (n=62)**	**Test-retest (n=65)**	**Sensitivity to change (n=44)**
Intentional self-harm	.82 (.70 – .89)	.87 (.78 – .92)	.79 (.62 – .89)
Unintentional self-harm	.40 (.00 – .64)	.91 (.86 – .95)	.87 (.77 – .93)
Risk from others	.59 (.32 – .75)	.89 (.83 – .94)	.79 (.61 – .88)
Risk to others	.32 (-.13 – .59)	.86 (.77 – .91)	.86 (.75 – .92)
Survival needs / disabilities	.45 (.08 – .67)	.86 (.77 – .91)	.71 (.48 – .84)
Psychological needs / disabilities	.40 (.00 – .64)	.77 (.62 – .86)	.76 (.56 – .87)
Social needs / disabilities	-.05 (-.37 – .75)	.77 (.63 – .86)	.54 (.17 – .75)
Total	**.58 (.30 – .75)**	**.87 (.79 – .92)**	**.80 (.63 - .89)**